EARLY GRADE READING ASSESSMENT BASELINE REPORT

KHYBER PAKHTUNKHWA PROVINCE

SEPTEMBER 2014

This publication was produced for review by the United States Agency for International Development by. It was prepared by Management Systems International (MSI) with School-to-School International (STS) under the Monitoring and Evaluation Program (MEP).

EARLY GRADE READING ASSESSMENT BASELINE REPORT

KHYBER PAKHTUNKHWA PROVINCE

Contracted under Order No. AID-391-C-13-00005

Monitoring and Evaluation Program (MEP)

DISCLAIMER

This study/report is made possible by the support of the American people through the United States Agency for International Development (USAID). The contents are the sole responsibility of Management Systems International and do not necessarily reflect the views of USAID or the United States Government.

ACKNOWLEDGEMENTS

We would like to thank the Education team of USAID/Pakistan for their forward planning to be able to collect baseline data before the roll out of the two important reading programs. Their support and responsiveness under a demanding timeline made this study possible. We would also like to thank the Elementary and Secondary Education Department of Khyber Pakhtunkhwa for their support of this activity. Finally, this effort would not have been possible without the dedication of our field teams of quality control officers and our local data collection partner, the Basic Education for Awareness, Reforms, and Empowerment (BEFARe).

CONTENTS

List of Tables and Figures

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
BEFARe	Basic Education for Awareness, Reforms and Empowerment
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DOE	Education Department
EGRA	Early Grade Reading Assessment
F.A.	Fellow in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Fellow in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
M.Sc.	Master of Science
MSI	Management Systems International
MT	Master Trainers
NEMIS	National Education Management Information System
PRP	Pakistan Reading Project
PTA	Parent Teacher Association
PTC	Parent Teacher Council
P.T.C.	Primary Teaching (Grade 12) Certificate
PTSMC	Parent Teacher School Management Committee
QCO	Quality Control Officer
SMC	School Management Committees
SPSS	Statistical Package for the Social Sciences
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

In 2013, Management Systems International (MSI) and School-to-School International (STS) conducted a baseline reading assessment for primary school children prior to the launching of two USAID-funded projects: the Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP). PRP is targeting improved reading for 910,000 children in Azad Jammu and Kashmir (AJK), Balochistan, the Federally Administered Tribal Areas (FATA), Gilgit-Baltistan (GB), the Islamabad Capital Territory (ICT), Khyber Pakhtunkhwa (KP), and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and administrators, and 2) improved reading instruction for children in the primary grades.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. This report covers the baseline assessment in KP which took place in September 2013. In May 2013, GB, AJK, and ICT were part of Round 1 of the baseline data collection, while data collection in KP, along with Sindh, was completed in September 2013 and data collection in Punjab, Balochistan, and FATA was completed in October 2013. The following activities were carried out for all of the provinces, including KP: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

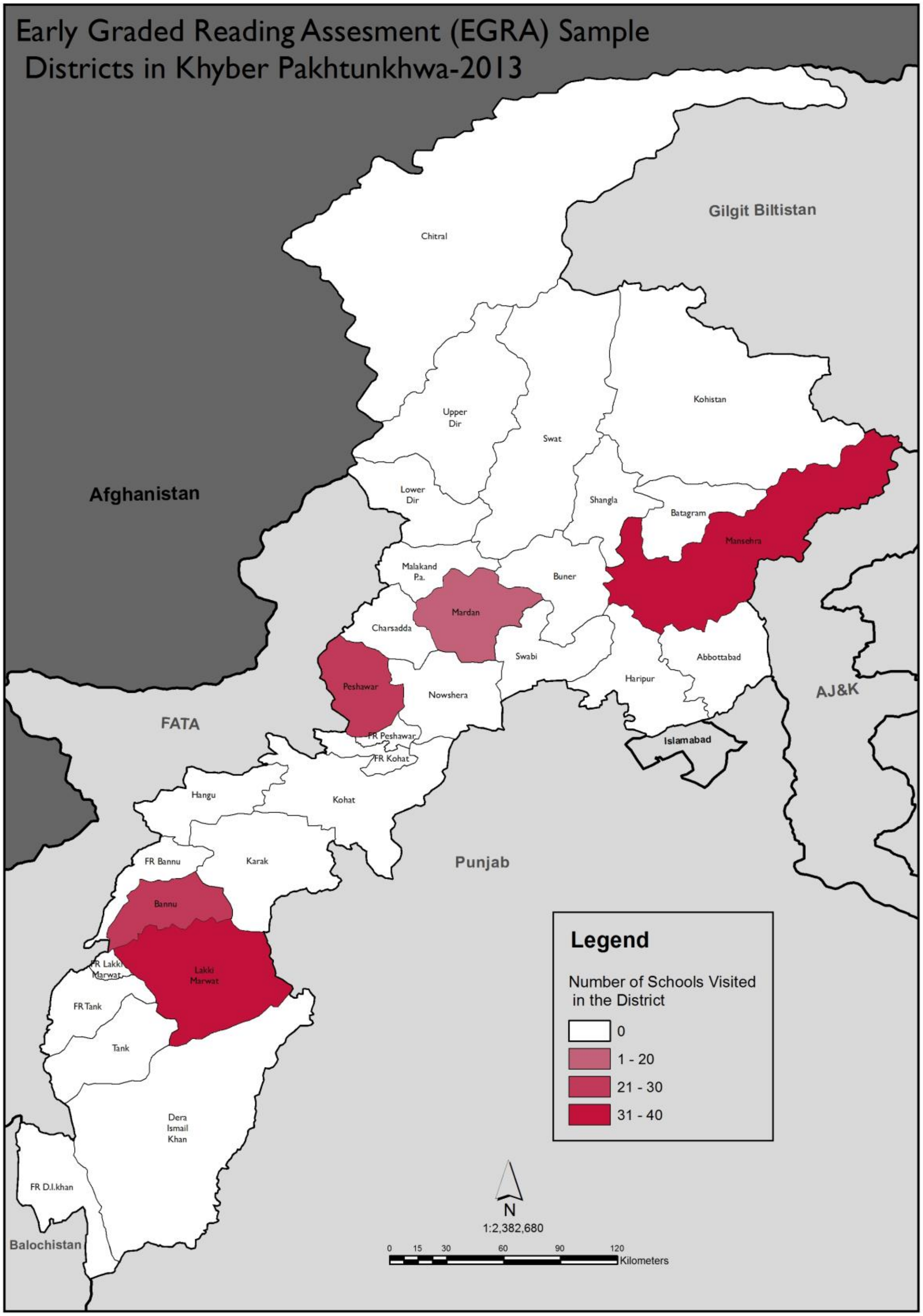
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: full treatment and light treatment. The full treatment group will receive both the first and second kinds of support, i.e., 1) policy, laws, and guidelines, and 2) improved instruction. The light treatment group will only receive the first kind of support.

In accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – will be assessed at three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), will be individually administered to over 30,000 children in 1,120 schools throughout the country. Over the course of the projects, the evaluators will compare the baseline results with those at the midline and endline to examine success in improving children's reading levels in Pakistan. The sampling was designed so that each province could be evaluated independently.

The long-term goal of this evaluation is to compare each province's baseline results to its midline and endline results, rather than other province's results. There are too many confounding variables – languages, curricula, administration dates, etc., that could render these comparisons meaningless. Furthermore, the evaluation is designed to investigate reading performance of the full and light treatment groups across time: baseline, midline, and endline. The differences between treatments will be fully investigated later, given the baseline data as the starting point for comparisons; the databases are designed for these comparisons at the follow-up stages. Therefore, in-depth comparisons between the full and light treatment groups are not useful at this time. Such comparisons at baseline could add some bias by facilitating competition between the two groups that could compromise the validity of the evaluation.

For the baseline in KP, all activities were completed by the end of January 2013, including a draft report. The EGRA baseline results was presented and discussed at a consultative meeting in Islamabad in April 2014. Representatives from the provincial Ministry of Education, USAID, and the contractors (MSI and STS) will attend the consultation. Revisions will then be made to this report based on the discussions between the stakeholders.

Map of Sampled Districts



Key Points

Several key points from the EGRA baseline assessment in KP are highlighted below:

Implementation

1. The KP evaluation involves two kinds of comparisons: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline. Given the long-term design of this evaluation, this baseline report will not statistically test the differences between groups' initial reading performance because doing so may confound the study by facilitating comparisons between the groups. This will however, present the baseline scores for each group. (Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.)

2. District selection into full and light treatment groups was finalized following consultative meetings between the Education Department (DOE) and USAID in February 2013.

3. EGRA in Urdu was used in the KP province. The EGRA tools, which have been administered in various forms in over 40 countries, were successfully adapted for use in Pakistan. These included individually administered reading tests for students, along with questionnaires for students, teachers, and head teachers. The Urdu version of the tools was piloted in AJK, ICT, and KP. The Sindhi version of the tools was piloted in Sindh.

4. A total of 140 schools, with 70 schools from each group (full and light treatment), were selected for the baseline.

5. The baseline data were collected in schools in a simple random sample of five districts. Schools in the Bannu (40) and Lakki Marwat (30) districts were chosen to represent the light treatment group, while schools in Mansehra (30), Mardan (20), and Peshawar (20) were sampled for the full treatment group.

6. For each district, a random sample of boys and girls schools was obtained, followed by a random sample of students in grades 3 and 5 within those schools. The results from this sample are presented in this report as a generalized view of the reading levels for students in the KP schools. Please note that district comparisons are not possible because the districts were not evenly sampled.

7. The target baseline sample for KP was 140 schools. The assessment tools were successfully administered in (with the percentage of the target reached in parentheses) 140 schools (100.0 percent) to 3,916 students (93.2 percent), 239 teachers (85.4 percent), and 140 head teachers (100.0 percent). The percentage of teachers is low because some teachers taught both grades and were not counted in survey results, which were analyzed by grade level.

8. The EGRA testing window for KP was in September 2013, and all schools were reached during this time period.

9. The validity and reliability of the tools for both languages was acceptable. Validity was assured through the adaptation process, which involved 17 educationists from throughout the country who participated in a workshop in Islamabad. Reliability was assured through the high quality of the assessment tools and the standardized administration of the tools in KP. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.

10. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.

11. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed reading tasks (fluency or reading rates). These scores provide a comprehensive picture of student performance. Analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Results

1. EGRA was administered to 1,940 grade 3 students and 1,976 grade 5 students. The reliability estimates were high for both grades (alpha = 0.85 for grade 3 and 0.83 for grade 5), indicating that the items worked well in measuring reading constructs at each grade level.

2. The task and item statistics showed that the EGRA discriminates well between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower and low-middle section of the difficulty range while p-values for grade 5 were higher, covering the mid-lower half and middle parts of the spectrum. All task scores at grades 3 and 5 had item-total correlations equal to or greater than 0.20, indicating good quality for these tasks. (Complete item statistics are listed in Annex 1).

3. Both grade levels did relatively better on the orientation to print, passage reading, and familiar word reading. They also had relatively low scores in comprehension (passage and listening). In addition, grade 3 showed difficulty with phonics (letter sound knowledge, phonemic awareness, and non-word reading) while grade 5 recorded low scores in phonemic awareness.

4. There was substantial progression from grade 3 to grade 5 on the summary score (15 points) and on all of the tasks scores – the greatest gains were in familiar word, passage, and non-word reading. This progress was consistent across gender and treatment groups.

5. There were differences between boys and girls on the task and summary scores, but most of these differences were small. Girls had higher scores on all tasks except orientation to print. The girls' summary scores were 6.5 points greater at grade 3 and 9 points higher at grade 5.

6. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). In terms of fluency, passage reading increased by 33.8 words correct per minute from grade 3 (35.5) to grade 5 (69.3). Although the passage was designed for grade 3, this difference shows that the reading levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically mastery of phonics, such as letter sound knowledge, phonemic awareness, and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in KP schools.

7. The average summary score for students from full treatment schools was about 3 points higher at grade 3 and 6.5 points higher at grade 5 than in light treatment schools (See Table 8). Most of the full treatment task scores were greater than the light treatment group, but these differences were small for most tasks. The largest differences were in the phonemic awareness and the two comprehension questions. These differences will be corrected statistically when progress for each group is measured during the midline and endline evaluations.

8. Student questionnaire findings revealed three interesting findings. The first positive finding was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes for both grades 3 and 5 students. Second, grade 3 summary scores increased with relative age (younger than normal, normal, older than normal age); older students in the grade had higher reading scores. However, by grade 5 that advantage was no longer significant. Third, KP

students are performing well on the Urdu test considering only 3 percent and 5 percent of the students in grades 3 and 5, respectively, reported speaking Urdu at home.

9. Student, teacher, and head teacher questionnaire findings were mostly inconclusive due to small sample sizes and the lack of variation in the scores that were related to their characteristics. For teachers, those who attended one or more in-service trainings had higher scores than those who never attended such trainings. For head teachers, attending one or two in-service trainings, along with in-service training in teacher reading, tended to relate to higher reading scores for grade 3 students. For the schools, the presence of a library and better infrastructure were associated with better student reading scores.

Evaluation Recommendations

Given the success of the baseline assessment in KP (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. This should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.

2. The EGRA items and tasks had good discrimination (quality) values and covered the low-to-middle part of the difficulty range. At baseline, the reading scores were relatively low for both grades and show room for growth. In addition, histograms and box pots provided evidence that the tool is expected to measure higher levels of reading-rate fluency that are anticipated following project-led interventions. Therefore, the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the methods at the school level.

4. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers (MTs), field supervisors, quality control officers (QCOs), and enumerators as used in the baseline.

5. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.

6. The analysis should follow the same procedures, with calculations of task scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be comparable so that improvements in students' reading can be accurately examined.

7. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.

8. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires for collecting data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER 1: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, Balochistan, FATA, GB, ICT, KP, and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and educational administrators, and 2) improved reading instruction for children in primary grades. Some districts in Pakistan will receive both kinds of support, i.e., "full treatment," while others will receive only the policy support, i.e., "light treatment." All schools within districts will receive the same type of treatment.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation baseline is taking place in 2013, prior to the launch of the reading interventions. In accordance with USAID program evaluation guidelines, samples of students in two selected grade levels – grade 3 and grade 5 – are being assessed throughout Pakistan so that independent baselines can be established in each province. Students at the same grade levels will be assessed at the midline and endline time points to evaluate the success of the interventions, taking into account the two treatment groups. The goal of the evaluation is to conduct a long-term assessment for both groups in each province.

This report covers KP province. Along with Sindh, KP was part the baseline data collection in September 2013; data from Pakistan's other six provinces were collected in May 2013 (ICT, AJK, GB) and October 2013 (Punjab, Balochistan, and FATA). The following activities were planned for all of the provinces, including KP:

1. Design – USAID required a cross-sectional design, i.e., assessing students at the same grade levels (grades 3 and 5) over the course of PRP and SRP. In most provinces, including KP, this was complemented by a quasi-experimental design with the two treatment groups (full and light).

2. Sampling – Schools were selected from the full and light treatment districts. The sample enabled the collection of student reading assessment data that were representative of the treatment groups, grade levels, and gender. There was also some stratification by urban/rural zones in Peshawar for the full treatment and Lakki Marwat for the light treatment group. Balance for this variable was not possible due to towns and districts within a group being primarily urban or primarily rural. Therefore, it was not appropriate to investigate the EGRA differences between urban and rural schools in the KP province.

3. Instrumentation – EGRA tools were developed, with tests at the grade 3 level in English, Sindhi, and Urdu, and questionnaires for teachers, head teachers and students in Urdu and Sindhi. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan. The Urdu instruments were used in KP.

4. Planning – A field administration plan was developed for the baseline administration that would ensure the reliability of the data collected. The plan specified the timeline, training, logistics, field activities, supervision, data entry, analysis, reporting, and quality control.

5. Training – Workshops were conducted to train all master trainers, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure clear comprehension and skills adequate to implement the EGRA tools.

6. Implementation – The baseline survey was implemented according to the plan. It ensured that all of the field activities took place in a standardized manner, as verified by the QCOs. The fieldwork was followed by data entry and preparation of a clean data file.

7. Analysis – Data were analyzed using spreadsheet (Excel) and statistical (SPSS) software. Experienced statisticians/psychometricians conducted the analysis, produced data tables and graphs, and ensured quality control.

8. Reporting – Provincial-level reports were produced and will be disseminated to the provincial education authorities. A template was developed according to guidelines from the USAID contract.

This report is organized into four chapters: 1) introduction, 2) methodology, 3) findings and results, and 4) conclusions and recommendations. Annexes with item statistics, box plots for the timed tasks, and a possible process for establishing a reading proficiency threshold follow the chapters.

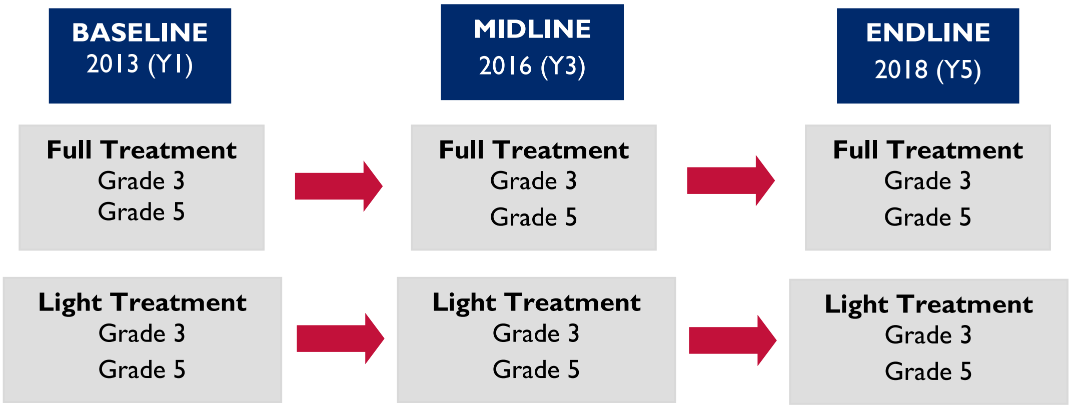
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the systems used for collecting the EGRA baseline data for schools in KP. There are sections on the evaluation design, timeline, sampling, instrument development, data collection, data entry, and data analysis.

Evaluation Design

Following USAID policy, a cross-sectional evaluation design was developed prior to the baseline data collection. As shown in Figure 1, the design features two grade levels (3 and 5) and three time points (baseline, midline, and endline). Different groups of grade 3 and grade 5 students will be compared against each other across the three time points. In the figure, the years for the midline and endline are approximate and may be altered in accordance with implementation of the PRP interventions.

FIGURE 1: EVALUATION DESIGN



Districts for the full treatment group were pre-selected by the DOE and USAID in KP. Since district-level selection for the two groups was not random, equivalence at baseline of the two treatment groups cannot be assured, and a quasi-experimental design was selected. In this design, any differences in scores at baseline (and midline and endline) will be statistically removed in the analysis, i.e., the two groups will be made statistically equivalent even though their average scores may be different. This will ensure fairness in the comparison of the full and light treatment groups. In addition, scores between the groups will not be statistically tested at baseline because the goal of the evaluation is to compare the long-term progress of both groups. Providing group comparisons at baseline may introduce potential competition between the groups and invalidate the experimental design. Group comparisons will be fully investigated at the endline assessment.

For the baseline assessment in KP, a random selection from the full treatment districts as selected for the PRP interventions resulted in the choice of Mansehra, Mardan, and Peshawar. Bannu and Lakki Marwat were randomly selected for the light intervention districts. For each treatment group and district, equal numbers of boys and girls schools were sampled for the EGRA testing. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

In KP, students were tested in Urdu, their main language of instruction. Some of the schools in the province use other languages during instruction (e.g., Pashtun), though their materials (e.g., textbooks) are in Urdu.

Timeline

The KP baseline, like the other provinces, was conducted according to a timeline that started in January and ended in January 2014, with draft submissions of reports to USAID in January 2014. The reports may then be distributed to the DOE and other stakeholders as appropriate (see Table 1 below).

The process began with the planning and design of activities, including creating preliminary sampling designs, selecting model EGRA tasks, recruiting staff, and budgeting/contracting. From February to August, the EGRA team, with participation from KP and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the district/school sampling. The data collection in KP took place in September 2013 followed by the data entry, analysis, and reporting in November and December. Presentations to the DOE and USAID were concluded in April 2014. The final report for KP was submitted in May 2014.

TABLE 1: TIMELINE (JANUARY 2013 TO MAY 2014)

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Plan and design EGRA activities	X	X															
Debrief to USAID and the DOE	X	X	X				X	X									
Prepare EGRA tools		X	X														
Prepare test administration manuals			X														
Train master trainers and enumerators								X									
Select and verify sample schools							X	X									
Administer EGRA									X								
Enter data										X							
Analyze baseline data										X	X						
Produce draft reports												X					
Produce presentations													X	X			
Disseminate draft reports												X	X				
Make presentations																X	
Revise and finalize reports																X	
Submit final reports to USAID																X	X

Sampling

The sampling for KP was finalized in July 2013 following meetings with USAID. The EGRA team conducted the school sampling in July. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample. As mentioned above, the schools in Mansehra, Mardan, and Peshawar implemented the full treatment, and Lakki Marwat and Bannu implemented the light treatment. The sampling for KP, as detailed in the sampling report for USAID,[1] is briefly summarized in the following sub-sections of this report.

Sampling Requirements

Since the minimum requirement was 15 students per grade level in grades 3 and 5, only schools meeting that requirement were eligible for sampling. Within the treatment groups (full and light), equal numbers of boys and girls schools (35 each) were selected.

Sampling Process and Field Verification

Due to the security concerns and other issues, not all districts in KP were considered for the PRP interventions or for the baseline assessment. From the chosen full treatment and light treatment districts, districts for the full and light treatment assessment groups were randomly selected. This resulted in a clustered sample. For the 35 boys and 35 girls schools in both the full and light treatment groups, the samples were divided among the selected districts according to the proportions of schools within those districts (stratified random sampling). An equal number of boys and girls schools were chosen within each group. For both groups, a second stratification was done at the "location" level, where schools were allocated by rural and urban. Table 2 shows the number of schools and replacement schools for both treatment groups per gender and location. Note that mixed schools may have been selected for some replacement schools due to not having enough options for replacement schools of strictly one gender. However, only students from the respective genders were included in those samples (i.e. if a mixed school was selected to replace a female school, only females were sampled).

[1] MSI (2013). *Pakistan EGRA Sampling Report.*

TABLE 2: SCHOOLS BY DISTRICT, TREATMENT, LOCATION, AND GENDER

District	Location	Schools	Percentage	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Full Treatment Group							
Mansehra	Rural	2,457	57	20	20	6	6
Mansehra	Urban	27	0	0	0	0	0
Mardan	Rural	417	10	4	4	1	1
Mardan	Urban	61	1	0	0	0	0
Peshawar	Rural	973	22	7	7	2	2
Peshawar	Urban	412	10	4	4	1	1
Sub-Total		**4,347**	**100**	**35**	**35**	**10**	**10**
Light Treatment Group							
Bannu	Rural	1,435	55	19	19	6	6
Bannu	Urban	59	2	1	1	0	0
Lakki Marwat	Rural	1,038	40	14	14	4	4
Lakki Marwat	Urban	62	3	1	1	0	0
Sub-Total		**2,594**	**100**	**35**	**35**	**10**	**10**
Total (both groups)		**6,941**	**100**	**70**	**70**	**20**	**20**

Once the schools were sampled, the QCOs, supplemented by EGRA senior managers, verified the samples in the field. This step was necessary due to two factors: 1) some inaccuracies in the National Education Management Information System (NEMIS) data, and 2) changes in student numbers since the time period when the schools had submitted their data to NEMIS. If the original schools had fewer than 15 students in either grade 3 or 5, a replacement school was selected and verified. At times, schools were retained if their student numbers were near the minimum.

Intended and Actual Samples

For the full treatment group, five schools – two boys and three girls schools – were substituted with schools randomly selected from the "replacement schools" list. The schools were replaced due to lower than expected numbers of students in the original samples. This was the same for the light treatment group where six schools – one boys and five girls schools – were replaced for the same reason. The actual numbers of students, teachers, and head teachers in the survey are presented in the results section.

Instrument Development

A brief summary of the instrument development process is presented below. The full results from the trans-adaptation, which involved educationists from KP, were presented in a report to USAID.[2]

[2] MSI (2013) *Pakistan EGRA Tools Trans-Adaptation Workshop Report.* June (Revised).

Trans-adaptation

In February 2013, the EGRA team used tasks from recent EGRA administrations in other countries to develop a model test. Led by two international and two national assessment specialists, the EGRA team then organized a trans-adaptation workshop in Islamabad. A total of 17 English, Sindhi, and Urdu language specialists from the DOEs and teacher training institutes throughout Pakistan – including one specialist from KP – participated in the workshop.

The trans-adaptation process involved the following with the local experts:

1. Discuss and choose reading tasks that would be of value to the baseline assessment in Pakistan;
2. Adapt each reading task using appropriate content in English, Urdu, and Sindhi; and
3. Ensure that the content would be suitable for grades 3 and 5 students.

The workshop resulted in a pilot EGRA test and pilot student, teacher, and head teacher questionnaires. The head teacher questionnaires included items about school characteristics.

Piloting

In March 2103, the EGRA English and Urdu tools were piloted in selected schools in AJK, ICT, and KP provinces while the Sindhi tools were piloted in June in Sindh. Four tools were included in the pilot: 1) a student response booklet (including the student questionnaire), 2) a student stimuli booklet, 3) a teacher questionnaire, and 4) a head teacher questionnaire. The EGRA team conducted the pilot sampling, trained the enumerators, arranged the logistics, and supervised the piloting. The team then entered the pilot data into a database, analyzed the data, and developed preliminary recommendations for final tools in preparation for the revision workshop. They also prepared a piloting report for USAID.[3]

Revision and Finalization

The EGRA team held a revision workshop in March for the Urdu and English tools with a limited number of experts from the trans-adaptation workshop. The Sindhi tools were revised in July with Sindhi language experts. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The team then finalized the four instruments for each language and submitted them to USAID. USAID made suggestions, particularly around the inclusion of reading- and library-related items into the questionnaires, which would provide information for the PRP and SRP. The English and Urdu instruments were approved and then used in the training workshops in advance of Round 1 data collection in May and Round 2 data collection in September for the KP tools. The final instruments were comprised of the following:

- Students: 16 informational items, 8 tasks (one with 2 sub-tasks), and 34 questionnaire items.
- Teachers: 15 informational items and 52 questionnaire items.
- Head teachers: 17 informational items and 37 questionnaire items.

These instruments are available for use by education officials.

[3] MSI (2013). *Pakistan EGRA Instrument Development and Pilot Data Analysis*.

Data Collection

Subcontractor Selection

The EGRA team, with the participation of USAID, issued a request for proposals and followed a set of criteria to select local subcontractors for the field data collection and data entry. In August, the Basic Education for Awareness, Reforms, and Empowerment (BEFARe) was chosen for both activities (data collection and data entry). MSI, STS, and BEFARe collaborated on the data collection in KP.

Data Collection

In August, EGRA senior managers trained MTs and QCOs during a two-week session in Islamabad. The MTs then spent one week, in Peshawar, training the BEFARe data collection team, which was comprised of one regional coordinator, two field supervisors, and 64 enumerators. The QCOs, coordinators, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations, the QCOs and field supervisors conducted a four-day refresher course for the enumerators in each district just prior to commencing data collection in the schools.

Over a 10-day period in September, the enumerators spent a day in each of the 140 schools to collect the baseline data in KP. The enumerators were in regular communication with the EGRA senior manager, QCOs, coordinator, and field supervisors to check on the status of data collection and to troubleshoot any issues. After collecting the data from the schools, the enumerators submitted their booklets to the supervisors and QCOs for verification and feedback. At the end of data collection, all booklets were returned to Islamabad for data entry.

Data Entry

Data Entry

In May 2013, the EGRA team developed a customized data entry application so that 1) the exact data from the booklets and questionnaires could be entered into a database, and 2) the computers used for data entry could be networked with a server. In September, the team trained the BEFARe data coordinator, four supervisors, and 36 data entry operators. In October and November, the EGRA and BEFARe teams entered the data for over 21,000 student booklets, along with questionnaires for the teachers and head teachers (Rounds 2 and 3). This total included approximately 4,200 booklets and questionnaires for KP.

Data Cleaning

In October and November, the EGRA and BEFARe teams conducted the data verification and reconciliation. Following USAID requirements, 100 percent of the data were entered twice (double data entry) and any discrepancies between the first and second databases were reconciled. A clean data file was then provided to the data analysis team.

Data Analysis

Methodology

In June, the EGRA statisticians and psychometrician developed a research plan that included the following steps: 1) reliability estimates, 2) task and item statistics, 3) mean and grand mean scores (percent correct scores), 4) data plots, 5) timed (fluency) and untimed task scores, and 6) questionnaire results. They used SPSS for the analysis. Some of the analyses were replicated to ensure that the calculations were accurate.

Descriptive analyses and inferential statistical comparisons were conducted by grade level and gender for the student scores, and for the three sets of questionnaire data.

Validity and Reliability

Validity evidence for the tests was derived from previous experiences with EGRA in other developing countries, as well as through the trans-adaptation process in Pakistan. The test developers targeted grade 3 for the level of the tasks. An assumption was that the grade 5 students should perform better than the grade 3 students on each of the tasks.

For reliability, a generally accepted method is to estimate the internal consistency reliability (Coefficient Alpha) of the test. The minimum reliability threshold is approximately 0.75 to 0.80 for tests of this nature. Reliability was estimated for each province. Table 3 shows the reliability estimates for grades 3 (0.85) and 5 (0.83) in KP. These reliabilities are excellent and lend credibility to the tests' internal consistency, indicating that the items are generally measuring similar reading constructs for both tests.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	n-count	Alpha
Urdu	Grade 3	9	1,940	0.85
	Grade 5	9	1,976	0.83

Note that there were actually eight tasks, but one of the tasks (Task 7) was administered and scored in two parts, so the equivalent of nine tasks were used for the analysis.

Score Calculation

The EGRA data was analyzed three ways. First, p-values and item-total correlations were generated for assessing the difficulty and discrimination of the items and tasks. Second, the percent correct for each task provided an indication of the KP students' mastery of the tasks, and third, KP students' fluency was assessed.

Item P-values and Item-Total Correlations

P-values and item-total correlations are classical test theory statistics that are used to evaluate the performance of individual items and the tasks they comprise. Item difficulty is measured by p-values, which range from 0.00 to 1.00. Higher p-values indicate easier items, because a higher percentage of students posted correct responses. The other classical statistic is the item-total correlation, and it ranges from -1.00 to +1.00. This statistic measures how close the item or task relates to the overall percent correct on the summary score. Values above 0.2 are an indication of a good item or task.

Percent Correct

The results of the EGRA testing were calculated using task and summary scores. Table 4 lists the tasks, stimuli, raw score ranges, and the method for calculating the task and summary scores on the test. For each of the tasks, the stimuli (items) (i.e., questions, letters, sounds, words, and non-words) were worth one score point. The score points were added, and since the range of raw scores varies across the tasks, the percent of correct scores was used to report all results. No weighting was used with the tasks to calculate the summary scores. Each task summary score was calculated using the total number correct and dividing it by the number of items. The overall Reading Summary Score was calculated by adding all of the task summary scores and dividing by nine (total number of tasks) to arrive at the average.

Timed Tasks Scores

The scores on the timed tasks were calculated by taking the number of correct responses times 60 seconds then dividing that number by the number of seconds used to read the stimulus. For instance, if a student read 75 letters correctly in 30 seconds, their letters-correct-per-minute score would be 150 (75 words x 60 seconds/30 seconds). Given another example, if a student read 50 words correctly in 30 seconds, his or her timed task score would be 100 words per minute (50 words x 60 seconds/30 seconds). Table 4 lists the number of stimuli per task. Recall the percent correct scores ranged from zero to 100. The method for calculating phonics and fluency scores yielded much higher maximum values, upwards of 200 at baseline (see task box plots in Annex 2, Figures A1 and A2).

TABLE 4: EGRA SCORE RANGES AND CALCULATIONS

Task (Subtest)	Stimuli	Score Range	Calculation
1. Orientation to print	5 questions (untimed)	0-5	Percent correct of answers
2. Letter name recognition	100 letters (timed)	0-100	Percent correct of letters
3. Phonemic awareness	10 questions (untimed)	0-10	Percent correct of words
4. Letter sound knowledge	100 sounds (timed)	0-100	Percent correct of sounds
5. Familiar word reading	50 words (timed)	0-50	Percent correct of words
6. Non-word reading	50 non-words (timed)	0-50	Percent correct of non-words
7a. Passage reading	60 words (timed)	0-60	Percent correct of words
7b. Passage comprehension	5 questions (untimed)	0-5	Percent correct of answers
8. Listening comprehension	3 questions (untimed)	0-3	Percent correct of answers
Reading Summary Score	-	-	Average of percent correct

An example of percent correct scores for each of the tasks and as a summary score is provided below. The raw score is divided by the maximum score (the highest score possible in the score range) to produce the percent correct score for each task. Then, the task scores are averaged to produce the summary score. Note that each of the task percent correct scores is weighted equally to provide the summary score.

TABLE 5: EXAMPLE OF EGRA PERCENT CORRECT AND SUMMARY SCORES

Task (Subtest)	Maximum Score	Raw Score	% Correct Score
1. Orientation to print	5	3	60.0%
2. Letter name recognition	100	68	68.0%
3. Phonemic awareness	10	5	50.0%
4. Letter sound knowledge	100	42	42.0%
5. Familiar word reading	50	34	68.0%
6. Non-word reading	50	25	50.0%
7a. Passage reading	60	50	83.3%
7b. Passage comprehension	5	2	40.0%
8. Listening comprehension	3	1	33.3%
Reading Summary Score	--	--	55.0%

An example of timed task scores (adjusted) is provided below for the five fluency tasks. The formula explained above is used (timed task score = raw score x 60 seconds/seconds used).

TABLE 6: EXAMPLE OF EGRA TIMED TASK SCORES

Task (Subtest)	Raw Score	Seconds Used	Timed Task Score
2. Letter name recognition	68	48	85.0
4. Letter sound knowledge	42	60	42.0
5. Familiar word reading	34	48	42.5
6. Non-word reading	25	40	37.5
7a. Passage reading	50	40	75.0

CHAPTER 3: FINDINGS AND RESULTS

This chapter presents the findings and results from the EGRA baseline in KP. There are sections on the student sample, task and item statistics, score calculation, task and summary scores, fluency task scores, and questionnaire findings.

Student Sample

The intended sample was 70 full and 70 light treatment schools. Within these schools, the target was to assess 15 students in each grade, totaling 4,200 students; 2,100 for each gender, for each treatment group, and per grade. Table 7 shows the number of students in the sample by grade and gender. For the full treatment group in grades 3 and 5, the actual samples were 93.9 and 93.4 percent of the intended sample, respectively. For the light treatment group the actual sample size was 90.4 percent for grade 3 and 93.6 percent for grade 5. The entire grade 3 sample was 92.1 percent and 93.5 percent for grade 5. The boys' percent (87.8) was lower than the girls' (97.9).

A small number of students in grade 3 (n = 5) and grade 5 (n = 12) did not complete the gender item on the questionnaire. Due to the missing gender codes, when analyzing the students by this characteristic the sample was 3,899 students, 92.8 percent of the intended 4,200 sample records. However, when the data were not analyzed by gender the total actual sample was 3,916 (93.2 percent of the intended 4,200 students).

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Treatment	Grade Level	Sample	Boys	Girls	Missing	Total
Full Treatment	Grade 3	Students	434	552	3	989
		% of Target	82.7%	105.1%	--	94.1%
	Grade 5	Students	448	533	2	983
		% of Target	85.3%	101.5%	--	94.0%
	Total	Students	882	1085	5	1,972
		% of Target	84.0%	103.3%	--	94.0%
Light Treatment	Grade 3	Students	474	475	2	951
		% of Target	90.3%	90.5%	--	90.7%
	Grade 5	Students	488	495	10	993
		% of Target	93.0%	94.3%	--	94.2%
	Total	Students	962	970	12	1,944
		% of Target	91.6%	92.4%	--	92.4%
Full and Light Treatment	Grade 3	Students	908	1027	5	1,940
		% of Target	86.5%	97.8%	--	92.4%
	Grade 5	Students	936	1,028	12	1,976
		% of Target	89.1%	97.9%	--	94.1%
	Total	Students	1,844	2,055	17	3,916
		% of Target	87.8%	97.9%	--	93.2%

Task and Item Statistics

Table 8 shows the statistics for the tasks for the KP sample. Two classical statistics are provided: p-values and item-total correlations. P-values indicate the average score of the students on the tasks, or the difficulty of the tasks for the students. The item-total correlations in the table are actually task-total correlations, which indicate the degree to which the tasks can discriminate between low- and high-achieving students; this is an indicator of the quality of the items. P-values can range from 0.00 to 1.00, with higher values indicating easier items. Item-total correlations can range from -1.00 to +1.00, with values above +0.20 indicating that the item (or task) is of good quality.

In Table 8 below, the task p-values for grade 3 ranged from 0.13 to 0.55, thus providing a spread on the lower and low-middle section of the difficulty spectrum. The p-values for grade 5 were higher, ranging from 0.27 to 0.72, or in the mid-lower half and middle parts of the range. All of the task scores in grades 3 and 5 had item-total correlations equal to or greater than 0.20, indicating good quality for these tasks. Complete item statistics are provided in Annex 1 at the end of this report.

TABLE 8: TASKS STATISTICS (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print	0.55	0.20	0.59	0.20
2. Letter name recognition	0.33	0.57	0.46	0.45
3. Phonemic awareness	0.34	0.31	0.44	0.36
4. Letter sound knowledge	0.19	0.55	0.27	0.41
5. Familiar word reading	0.44	0.82	0.71	0.75
6. Non-word reading	0.25	0.76	0.45	0.71
7a. Passage reading	0.46	0.82	0.72	0.74
7b. Passage comprehension	0.13	0.62	0.27	0.61
8. Listening comprehension	0.17	0.42	0.30	0.49

Task and Summary Scores

The next part of the analysis involved plotting the summary scores. Histograms of the summary scores (Figures 2 and 3) show that the distributions are moving from left to right from grade 3 to grade 5, which is strong evidence that the children are learning basic skills at the primary school level. In addition, as with the task and item statistics, it also shows that there is room for growth at each grade level. The main goal of the intervention is to see movement of the score distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

FIGURE 2: GRADE 3 SUMMARY SCORES

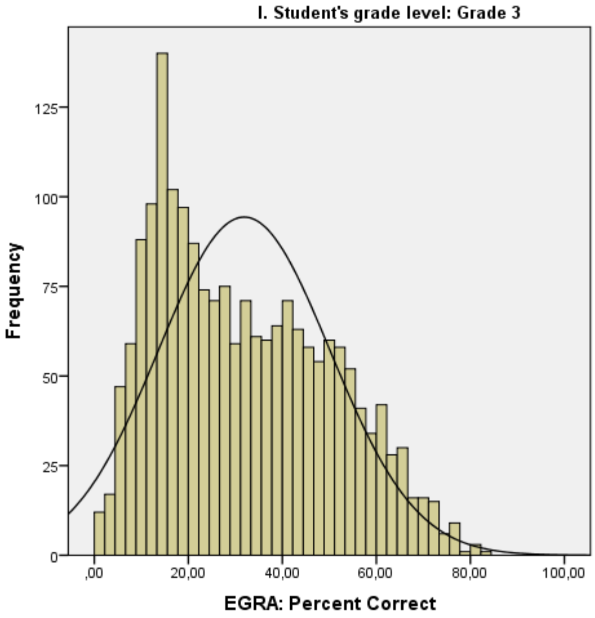
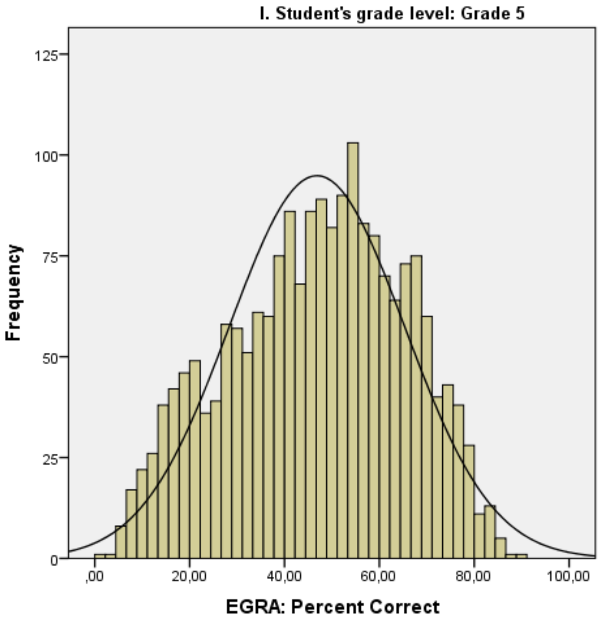


FIGURE 3: GRADE 5 SUMMARY SCORES



Tables 9 and 10, and Figures 4 and 5, provide the average scores by task using percent correct scores. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in KP was 54.9 percent for grade 3 and 59.4 percent for grade 5. The same type of calculation was made for each student and each task. The table also includes the differences from grade 3 to grade 5, e.g., 59.4 percent minus 54.9 percent equals 4.5 percentage points.

Both grade levels did relatively better on the orientation to print, passage reading, and familiar word reading. They also had relatively low scores in letter sound knowledge and in comprehension (passage and listening). In addition, grade 3 showed difficulty with letter name recognition and phonics (phonemic awareness and non-word reading), while grade 5 recorded low scores in phonemic awareness.

There was also substantial progression from grade 3 to grade 5 on the summary score (15 points). The greatest gains were in familiar word reading (26 points), passage reading (26 points), and non-word reading (20 points). This gain was consistent across treatment groups - 17 and 13 percentage points for the full and light treatments groups, respectively. Girls had a 16-point gain and the boys had a 14-point increase. In areas where there are large differences, interventions at grade 3 could have particularly large effects in accelerating children's learning.

TABLE 9: PERCENT CORRECT SCORES BY GRADE AND TASK (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3	Grade 5	Difference (G5 – G3)
1. Orientation to print	54.9%	59.4%	4.5% points
2. Letter name recognition	33.0%	45.6%	12.6% points
3. Phonemic awareness	34.1%	44.4%	10.3% points
4. Letter sound knowledge	19.4%	27.2%	7.8% points
5. Familiar word reading	44.3%	70.5%	26.2% points
6. Non-word reading	25.2%	44.9%	18.3% points
7a. Passage reading	46.2%	72.1%	25.9% points
7b. Passage comprehension	12.8%	27.4%	14.6% points
8. Listening comprehension	16.8%	30.1%	13.3% points
Reading Summary Score	31.9%	46.8%	14.9% points

For both grades, the full treatment scores were higher for most tasks, except for orientation to print and letter sound knowledge. The largest discrepancy was in passage and listening comprehension for both grades. The full treatment group summary score was 10 points higher for grades 3 and 5, respectively. This discrepancy will be corrected statistically at the midline and endline. Because this is a baseline report, the group differences will not be statistically tested at this time.

TABLE 10: PERCENT CORRECT SCORES BY GRADE, TASK, AND GROUP

Task (Subtest)	Full		Light	
	Grade 3	Grade 5	Grade 3	Grade 5
1. Orientation to print	50.3%	58.2%	59.7%	60.4%
2. Letter name recognition	33.4%	45.9%	32.6%	45.2%
3. Phonemic awareness	36.1%	49.5%	32.0%	39.3%
4. Letter sound knowledge	19.2%	25.7%	19.7%	28.7%
5. Familiar word reading	45.6%	72.8%	43.1%	68.3%
6. Non-word reading	26.6%	48.4%	23.8%	41.4%
7a. Passage reading	48.1%	75.3%	44.2%	68.8%
7b. Passage comprehension	17.0%	36.7%	8.4%	18.1%
8. Listening comprehension	22.5%	38.4%	10.9%	21.9%
Reading Summary Score	33.2%	50.1%	30.5%	43.6%

FIGURE 4: FULL TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK

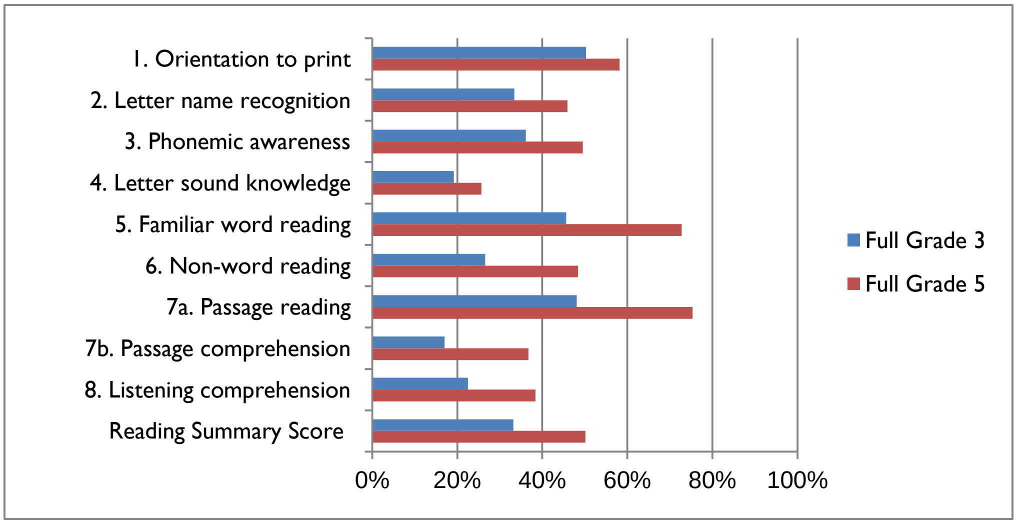
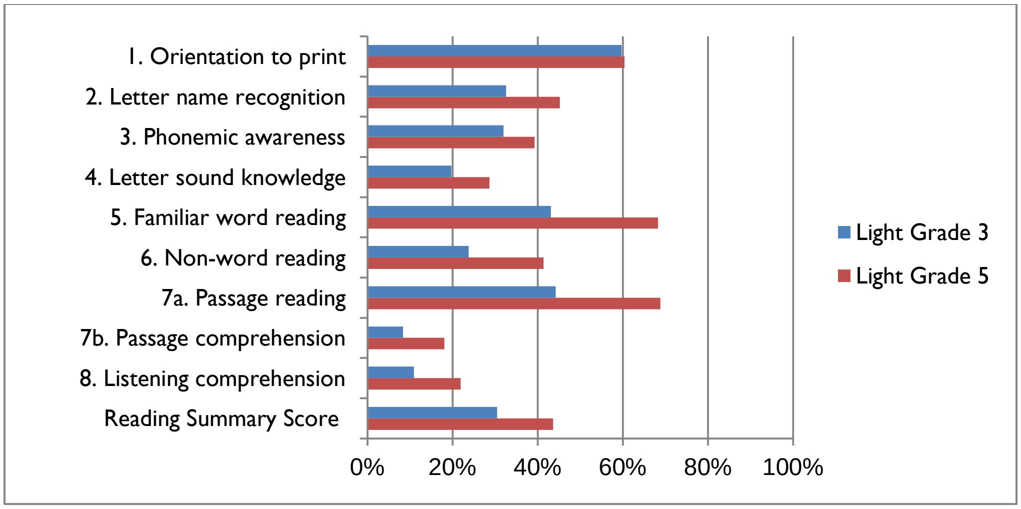


FIGURE 5: LIGHT TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK



The boys and girls had relatively higher scores in the same tasks: orientation to print, familiar word reading, and passage reading (Table 11 and Figures 6 and 7). In looking at the differences between the genders, girls had higher scores on all tasks except for orientation to print. One task in particular, passage comprehension, was difficult for grade 3 boys. The summary scores for girls were 6.5 points greater at grade 3, and 9.1 points higher at grade 5. The largest discrepancies between the gender groups were in familiar word reading, passage reading, and the two comprehension tasks.

TABLE 11: PERCENT CORRECT SCORES BY GRADE, TASK, AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	56.4%*	53.7%	62.7%*	56.2%
2. Letter name recognition	30.8%	34.9%*	42.4%	48.4%*
3. Phonemic awareness	32.9%	35.1%*	43.2%	45.3%
4. Letter sound knowledge	16.1%	22.4%*	24.4%	29.8%*
5. Familiar word reading	38.4%	49.6%*	61.8%	78.2%*
6. Non-word reading	22.9%	27.5%*	39.0%	50.3%*
7a. Passage reading	39.7%	52.2%*	63.0%	80.4%*
7b. Passage comprehension	7.3%	17.7%*	17.4%	36.4%*
8. Listening comprehension	12.2%	21.1%*	24.6%	35.2%*
Reading Summary Score	28.4%	34.9%*	42.0%	51.1%*

* Indicates that the performance of the group was significantly higher, p< 0.01

FIGURE 6: GRADE 3 PERCENT CORRECT SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

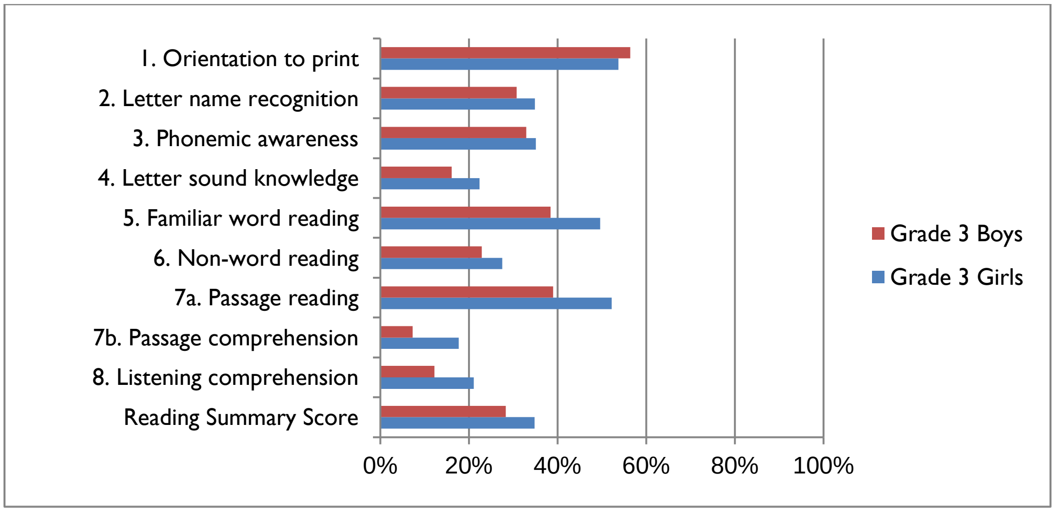
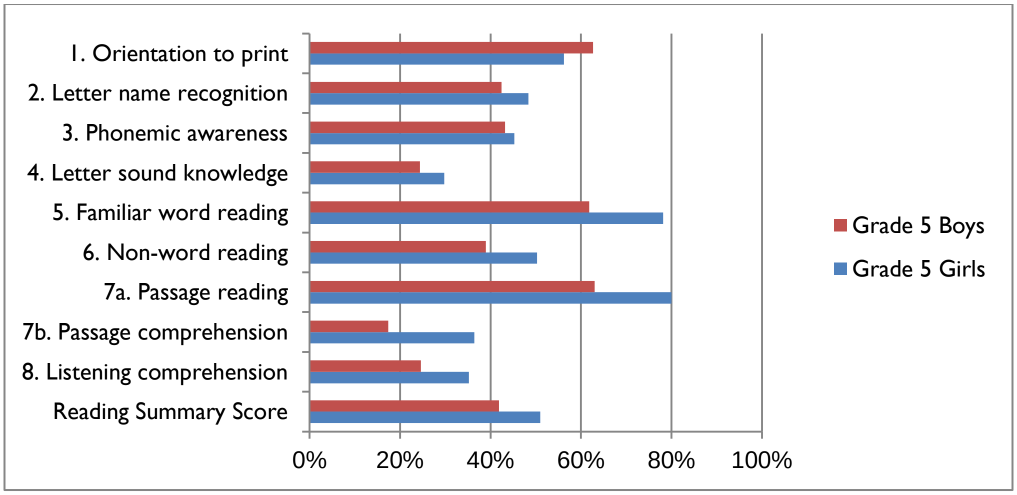


FIGURE 7: GRADE 5 PERCENT CORRECT SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)



The final table in this section (Table 12) further disaggregates the scores by treatment group, grade level, and gender. As seen in the tables above, the full treatment group scored higher on many of the tasks, which will be statistically corrected at the midline and endline.

There were some variations in the scores by gender and treatment group. For instance, on many of the tasks, the girls scored higher than the boys in the full treatment group, but the boys scored higher than the girls in the light treatment group. Further investigation would be required to determine the reasons for this trend.

TABLE 12: PERCENT CORRECT SCORES BY GROUP, GRADE, GENDER AND TASK

Task (Subtest)	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
1. Orientation to print	48.4	52.1	57.9	58.3	62.9	56.3	66.2	54.5
2. Letter name recognition	31.1	35.7	42.1	49.9	30.6	34.7	42.8	47.8
3. Phonemic awareness	34.3	38.0	46.0	52.9	32.5	31.4	40.9	37.6
4. Letter sound knowledge	16.8	21.4	22.4	29.0	14.4	25.3	25.5	32.2
5. Familiar word reading	42.5	48.7	65.3	80.1	35.3	51.3	59.9	76.8
6. Non-word reading	24.9	28.5	42.8	54.1	20.7	27.6	35.7	47.5
7a. Passage reading	43.0	51.5	66.5	81.7	36.2	52.0	59.2	76.7
7b. Passage comprehension	11.7	22.1	26.9	46.7	4.4	13.0	10.1	26.3
8. Listening comprehension	19.6	25.6	33.4	43.4	6.6	15.8	16.7	27.2
Reading Summary Score	30.1	36.0	44.9	55.2	26.9	34.1	39.6	47.4

Timed Tasks: Phonics and Reading-Rate Fluency Scores

Fluency is a measure of reading efficiency. On the Pakistan EGRA Baseline, there were two types of fluency measures: phonics and reading rate. The phonics-fluency subtest included letter name recognition, letter sound knowledge, and non-word reading, whereas, the reading-rate fluency subtest consisted of familiar word and passage reading.

Tables 14 to 17 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Table 11 has the maximum raw scores attained by students on each task at each grade level. Tables 14 to 17have mean scores for the students. In addition, adjustments were made to the raw scores for those students who finished the task before the end of one minute. For instance, if a student read 50 words correctly in 30 seconds, their words correct per minute score would be 100 (50 words x 60 seconds/30 seconds). Because these calculations are different from percent correct, the maximum scores are higher (see Figures A1 and A2 in Annex 2). Table 13 provides the baseline maximum scores at grade 3 and 5 for the five timed tasks.

TABLE 13: BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	223	158
4. Letter sound knowledge	117	166
6. Non-word reading	169	140
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 3**
5. Familiar word reading	142	177
7a. Passage reading	210	257

Both grades showed the same pattern in fluency (Table 14). Students in both grades performed best on passage reading, followed by familiar word reading, letter name recognition, letter sound knowledge, and lastly, non-word reading. The phonics tasks were the most difficult. The areas of greatest progress from grade 3 to grade 5 were the two reading-rate tasks: familiar word reading (19.4) and passage reading (33.8).

TABLE 14: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	34.8	47.2	12.4 points
4. Letter sound knowledge	32.4	40.1	7.7 points
6. Non-word reading	26.2	35.4	9.2 points
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 5**	**Difference (G5 – G3)**
5. Familiar word reading	37.9	57.3	19.4 points
7a. Passage reading	35.5	69.3	33.8 points

In comparing the treatment groups, there were minimal differences in the fluency tasks between the two groups (Table 15). For both grades, passage reading had the highest difference, favoring the full treatment group by 4.2 and 11.4 words correct per minute for grades 3 and 5, respectively. Again, these differences will be corrected statistically in the midline and endline evaluations.

TABLE 15: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GROUP

Phonics Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
2. Letter name recognition	34.3	35.3	46.8	47.7
4. Letter sound knowledge	31.4	33.6	40.0	40.2
6. Non-word reading	26.4	25.9	35.3	35.7
Reading-Rate Fluency Subtest	**Grade 3**		**Grade 5**	
	Full	**Light**	**Full**	**Light**
5. Familiar word reading	38.6	37.2	58.4	56.1
7a. Passage reading	37.6	33.4	75.0	63.6

Girls in both grades had significantly higher fluency scores than the boys on each of the timed reading tasks (Table 16). In grade 3, passage reading posted the greatest difference (11.9) favoring the girls. At grade 5 the gender disparity is more pronounced, for example, a 31.4 words-correct-per-minute discrepancy for passage reading and a 17.0 difference in familiar word reading.

TABLE 16: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	31.7	37.4	43.1	51.1
4. Letter sound knowledge	27.5	36.6	34.3	45.8
6. Non-word reading	25.5	26.5	30.1	40.0
Reading-Rate Fluency Subtest	**Grade 3**		**Grade 5**	
	Boys	**Girls**	**Boys**	**Girls**
5. Familiar word reading	33.5	41.4	47.8	65.3
7a. Passage reading	29.1	41.0	52.7	84.1

The final table in this section (Table 17) further disaggregates the scores by treatment group, grade level, and gender. As with the percent correct scores, the light treatment group scored higher on some of the tasks, which will be statistically corrected at the midline and endline.

TABLE 17: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GROUP, GRADE, AND GENDER

Phonics Fluency Subtest	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
2. Letter name recognition	31.0	35.8	42.3	49.6	31.1	34.9	42.9	48.8
4. Letter sound knowledge	17.9	20.2	23.1	24.8	14.4	25.2	25.7	32.5
6. Non-word reading	13.3	15.8	23.8	61.9	11.2	14.2	19.5	28.6

Reading-Rate Fluency Subtest	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
5. Familiar word reading	24.9	32.1	43.9	31.5	22.0	31.8	40.0	61.0
7a. Passage reading	31.5	42.1	58.1	88.8	26.9	39.8	47.8	79.1

Questionnaire Findings

Selected results are presented below, including for those characteristics or items that showed significant differences in student scores. Due to the students having the same language, the results were combined for the full and light treatment groups to increase the sample size and more accurately detect effects between the categories. Note that there were some students, teachers, and head teachers who did not respond to certain questionnaire items; they were labeled as missing. The total averages for the summary scores were calculated based on those who responded.

Since these are baseline data, reporting on the full and light treatment groups together will not affect the analyses at midline and endline. We combined the survey data for the groups since some of the questions led to reporting by relatively small categories (e.g., for teacher qualifications) and we wanted to know whether the survey results were associated with the student scores in general.

In addition, since the samples were by treatment group, the results will be generalized to the populations for each group. This will be done prior to the midline. The results will be generalized to by calculating sampling weights, applying the weights to the results, and then generalizing to the population by treatment group. We will also do this for the midline and endline. The current analyses only apply to the sampled districts.

Statistical significance was determined based on *t*-tests for indicators with two categories and analyses of variance for indicators with three or more categories (with post-hoc pairwise comparisons). The significance value was set at $p < 0.05$; a 95 percent confidence level.

Student Questionnaires

One survey question asked the students what language was spoken in the home. At grade 3, 94 percent spoke Pashto, 5 percent spoke Urdu, and less than 1 percent spoke Punjabi, Shina, or English. Similarly, at grade 5, 96 percent of the families conversed in Pashto and 3 percent in Urdu. Less than 1 percent of the grade 5 students spoke Punjabi or Shina at home. Urdu was not the primary language for most KP students.

Table 18 has summary scores by student age. According to the National Education Policy (2009), the official age of the students at the beginning of the different grade levels of primary education is 6 to 10 years old. Since the baseline took place during the school year, the normal ages for this analysis were set at 8 to 9 years old for grade 3 and 10 to 11 years old for grade 5. The students were placed into three categories: younger than normal age for their grade, normal age, and older than normal age. At grade 3, there were significant differences among all three age groups; in all comparisons, younger students had lower average summary scores. Conversely, at grade 5, the scores were not significantly different among age groups, eradicating the older-student advantage by grade 5.

TABLE 18: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	68	19.6%*	80	43.7%
Normal age	791	29.7%*	878	46.7%
Older than normal age	1,073	34.2%*	1,004	47.3%
Missing	8	-	8	
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p < 0.05 level

Table 19 shows the summary scores according to whether the student reads the Quran at home. There were significant differences in both grades favoring students who read the Quran.

TABLE 19: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	44	20.6%	35	36.1%
Yes	1,882	32.2%*	1,925	47.1%*
Missing	14	-	16	-
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

Table 20 depicts the differences in scores based on whether there is a library at the school. There was a significant difference for the grade 5 students, but not a significant difference for grade 3 students.

TABLE 20: SUMMARY SCORES BY THE PRESENCE OF A LIBRARY AT THE SCHOOL

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,018	32.2%	1,242	46.4%
Yes	695	33.1%	598	50.0%*
Missing	227	-	136	-
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

In Tables 21 to 23, the data showed that the existence of newspapers, magazines, and books generally made a difference in reading scores for both grades.

TABLE 21: SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,195	30.1%	1,060	44.3%
Yes	745	34.8%*	916	49.7%*
Missing	0	-	0	-
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 22: SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,827	31.5%	1,806	46.2%
Yes	113	37.3%*	170	53.7%*
Missing	0	--	0	--
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 23: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,601	31.3%	1,704	46.4%
Yes	339	34.5%*	272	49.5%*
Missing	0	--	0	--
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

The final set of student questions (in Tables 24 to 26 pertained to children's reading habits at home. In general, these habits made a difference in student scores in all cases for both grades. Having someone read to children at home, having children read to someone else at home, and children reading silently at home was related to higher reading scores.

TABLE 24: SUMMARY SCORES BY CHILDREN HAVING SOMEONE READ TO THEM AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	851	29.3%	908	44.5%
Yes	1,072	34.0%*	1,054	49.0%*
Missing	17	-	14	-
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 25: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	981	29.0%	907	44.4%
Yes	946	34.9%*	1,052	49.2%*
Missing	13		17	
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.001

TABLE 26: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	334	28.2%	256	43.3%
Yes	1,598	32.7%*	1,415	47.4%*
Missing	8		5	
Total	1,940	31.9%	1,976	46.8%

* Indicates that the performance of the group was significantly higher, p< 0.001

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to descriptive statistics, i.e., no group comparisons. Tables 27 and 28 provide information on teacher academic and professional qualifications, neither of which showed consistent patterns in the student scores.

TABLE 27: SUMMARY SCORES BY TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc./M.Phil.	41	31.6%	58	44.7%
B.A./B.Sc.	31	32.1%	25	50.2%
F.A./F.Sc.	22	36.1%	18	47.0%
Matric	28	31.5%	16	54.5%
Missing	0	--	0	--
Total	122	31.9%	117	46.8%

TABLE 28: SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	7	33.9%	21	43.7%
B.Ed.	37	31.4%	37	45.5%
C.T.	18	33.7%	13	42.7%
P.T.C.	60	32.7%	46	52.3%
Missing	0	--	0	--
Total	122	31.9%	117	46.8%

In an analysis of student scores by teacher age and experience, there were no consistent patterns of younger or older teachers, or those with less or more experience, relating to lower or higher student scores (Tables 29 and 30). Again, small teacher sample sizes made drawing conclusions difficult.

TABLE 29: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	65	33.7%	64	48.1%
Between 41 and 50	42	31.8%	39	45.8%
51 and more	11	27.3%	7	51.8%
Missing	4	--	7	--
Total	122	31.9%	117	46.8%

TABLE 30: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	47	32.9%	48	45.1%
Between 11 and 20	34	33.8%	38	49.0%
Between 21 and 30	32	30.8%	22	47.6%
31 or more	4	35.0%	6	54.8%
Missing	5	-	3	
Total	122	31.9%	117	46.8%

Teachers who attended one or more in-service trainings had higher scores than those who never attended such trainings (Table 31). Once more, any differences should be interpreted with caution due to the small sample size.

TABLE 31: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	68	29.9%	63	45.0%
One time	40	36.5%	41	50.5%
Two times	7	34.2%	7	48.1%
Three times	4	36.1%	4	47.6%
Missing	3	-	2	
Total	122	31.9%	117	46.8%

Head Teacher Questionnaires

Similar to the teachers, the sample size for the head teacher questionnaires was small, so data interpretations should be treated with caution. Tables 32 and 33 show reading scores by the head teachers' academic and professional qualifications. In general, the results show that higher teacher academic and professional qualifications show no definitive pattern in student scores.

TABLE 32: SUMMARY SCORES BY HEAD TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc./M.Phil.	41	31.9%	41	45.2%
B.A./B.Sc.	48	30.0%	48	46.6%
F.A./F.Sc.	47	32.4%	47	48.6%
Matric	3	48.5%	3	51.9%
Missing	1	-	1	-
Total	140	31.9%	140	46.8%

TABLE 33: SUMMARY SCORES BY HEAD TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	8	37.5%	8	52.8%
B.Ed.	30	31.3%	30	45.0%
C.T.	49	29.3%	49	45.6%
P.T.C.	51	33.6%	51	48.2%
Missing	2	-	2	-
Total	140	31.9%	140	46.8%

Tables 34 and 35 provide information on head teachers' experience and in-service training. For both grades, no discernible pattern was revealed in the head teachers' years of experience. In terms of in-services training, head teachers who attended one or more in-service trainings had higher grade 3 scores than those who did not attend the trainings, however, this observation was not found with grade 5 scores. Again, any differences should be interpreted with caution due to the small sample size.

TABLE 34: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	27	32.3%	27	44.4%
3 to 5	16	31.3%	16	49.3%
6 to 10	25	32.7%	25	45.4%
11 or more	36	29.9%	36	47.4%
Missing	36	-	36	-
Total	140	31.8%	140	46.7%

TABLE 35: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	95	30.6%	95	45.6%
1 time	31	35.5%	31	51.1%
2 times	9	25.6%	9	45.1%
More than 2 times	3	47.6%	3	52.2%
Missing	2	-	2	-
Total	140	31.9%	140	46.8%

Tables 36 and 37 provide data on head teachers' support to teachers in reading and the training that head teachers received in teaching reading. There were too few head teachers that reported not supporting teachers in reading (14); therefore the sample size is too small to make valid conclusions. However, there were slightly higher reading scores shown when the head teacher received training in teaching reading.

TABLE 36: SUMMARY SCORES BY HEAD TEACHER SUPPORT TO TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	14	32.7%	14	48.5%
Yes	124	31.8%	124	46.9%
Missing	2	-	2	
Total	140	31.9%	140	46.8%

TABLE 37: SUMMARY SCORES BY HEAD TEACHER TRAINING IN TEACHING READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	95	30.7%	95	45.6%
Yes	43	34.3%	43	50.0%
Missing	2	-	2	
Total	140	31.9%	140	46.8%

School Characteristics

The final section provides information on school characteristics (from the head teacher questionnaires) by student summary scores. As with the teacher and head teacher characteristics, most patterns appeared to be inconclusive (Tables 38 to 42). Urban schools had higher reading scores than those in rural settings. The girls schools performed better than the boys and mixed-gender schools. Impressively, all but three schools had parent-teacher organizations. Schools with libraries (only 27 percent) had higher scores at grade 3, but only slightly higher scores at grade 5. Lastly, better infrastructure seemed to be related to higher student reading scores.

TABLE 38: SUMMARY SCORES BY SCHOOL LOCATION

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Rural school	129	31.1%	129	46.5%
Urban school	11	38.3%	11	50.0%
Missing	0	--	0	--
Total	140	31.9%	140	46.8%

TABLE 39: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Boys school	61	29.3%	61	44.9%
Girls school	56	34.9%	56	51.0%
Mixed Gender	20	29.8%	20	41.3%
Missing	3	-	3	
Total	140	31.9%	140	46.8%

TABLE 40: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	3	31.3%	3	44.4%
Yes	136	31.8%	136	47.0%
Missing	1		1	
Total	140	31.9%	140	46.8%

TABLE 41: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	98	32.6%	98	47.0%
Yes	38	29.9%	38	47.1%
Missing	4		4	
Total	140	31.9%	140	46.8%

TABLE 42: SUMMARY SCORES BY INFRASTRUCTURE (DRINKING WATER, ELECTRICITY, TOILETS)

Number of Infrastructures (Water, Electricity, Toilets)	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	10	26.0%	10	41.5%
1	24	28.2%	24	48.8%
2	40	31.1%	40	45.5%
3	66	34.2%	66	47.7%
Missing	0	-	0	-
Total	140	31.9%	140	46.8%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions and recommendations from the KP EGRA baseline. The conclusions are organized according to the two main sections in the report: 1) design and methodology, and 2) findings and results. There are also recommendations based on the instrument development, data collection, data entry, and analysis.

Design and Methodology

1. The design followed USAID evaluation guidelines for a cross-sectional approach. This will allow for an examination of the progress of students in grades 3 and 5 over the life of the PRP. The province also has one instructional language, Urdu. (However, less than 5 percent of the students reported speaking Urdu in the home.) In addition, KP has two treatment groups: full and light. This will allow for an evaluation of the full treatment effects above and beyond those of the light treatment.

2. The sampling issues were addressed as well as could have been expected. In a limited number of schools, there was an issue of a lack of the requisite number of students per grade level. The actual sample of schools was 100 percent and the actual sample of students reached 93.2 percent of the intended sample.

3. The EGRA test in Urdu administered in KP was of good quality. The reliability estimates were in the high part of the range (greater than 0.80) of previous EGRA administrations in other countries. The task statistics were acceptable, with an appropriate range of p-values and item-total correlations that were at an acceptable level of quality. The characteristics of the tests were such that it should be a strong measure of potential progress over time due to project-led interventions. As with any test, there may be ways to improve on the task and item statistics for the midline and endline. The passage reading fluency scores were relatively high for grade 5. The evaluation team may want to investigate the appropriateness of the difficulty level of this task for the midline and endline assessments. However, changing the difficulty levels of the passage after baseline would require extensive work in equating the two passages in order to make valid comparisons from baseline to midline and endline scores.

4. The field implementation was successful, though there were difficulties to overcome, including the low actual enrollment of students in some schools. There was a high level of standardization reported by the quality control officers, which they attributed to the effective training process by the EGRA team. The team paid careful attention to detail in the logistics and test administration, which was reflected in the low error rates in the booklets and in the data entry.

Findings and Results

The KP evaluation involves two kinds of comparisons: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline. Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.

Several key findings emerged from the baseline assessment in KP. These are as follows:

1. EGRA was administered to a robust sample at each grade level (3 and 5) and in each group (full and light treatment). Test reliabilities were very good, showing that the EGRA tasks and items worked well in measuring reading constructs at both grade levels. The task and item statistics showed that EGRA discriminates well between low- and high-achieving students in both grades. They also showed that there is adequate room for growth by students in each grade level.

2. Both grade levels did relatively better on the orientation to print, passage reading, and familiar word reading. They had relatively low scores in comprehension (passage and listening). In addition, grade 3 showed difficulty with phonics (letter sound knowledge, phonemic awareness, and non-word reading), while grade 5 recorded low scores in phonemic awareness.

3. There was substantial progression from grade 3 to grade 5 on the summary score (15 points) and on all of the tasks scores – the greatest gains were in familiar word, passage, and non-word reading. This progress was consistent across gender and treatment groups.

4. There were differences between boys and girls on the task and summary scores, but most of these differences were small. Girls had higher scores on all tasks except orientation to print. The girls' summary scores were 6.5 points greater at grade 3 and 9.1 points higher at grade 5.

5. The average summary score for students from full treatment schools was about 3 points higher at grade 3 and 6.5 points higher at grade 5 than in light treatment schools (see Table 8). Most of the full treatment task scores were greater than the light group, but these differences were small for most tasks. The largest differences were in the phonemic awareness and the two comprehension questions. These differences will be corrected statistically when progress for each group is measured during the midline and endline evaluations.

6. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics fluency (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Passage reading increased 33.8 words correct per minute from grade 3 (35.5) to grade 5 (69.3). Although the passage was designed for grade 3, this difference shows that the reading levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically, mastery of phonics, such as letter sound knowledge, phonemic awareness, and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in KP schools.

7. The student questionnaires revealed three interesting findings. The first positive finding was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes for both grade 3 and 5 students. Second, grade 3 summary scores increased with relative age (younger than normal, normal, older than normal age); older students in the grade had higher reading scores. However, by grade 5 that advantage was no longer significant. Third, KP students are performing well on the Urdu test considering only 3 percent and 5 percent of the students in grades 3 and 5, respectively, reported speaking Urdu at home.

8. School, teacher and head teacher questionnaire findings were mostly inconclusive due to small sample sizes and the lack of variation in the scores that were related to their characteristics. Students had higher scores when their teachers attended one or more in-service trainings. For head teachers, attending one or two in-service trainings along with in-service training in teacher reading tended to relate to higher reading scores for grade 3 students. For the schools, the presence of a library and better infrastructure were associated with better student reading scores.

Evaluation Recommendations

1. The instrument development and trans-adaptation process was comprehensive and resulted in high quality EGRA tools. This should be repeated as soon as possible with the tasks that need to be changed for the midline and endline tools (to minimize test-retest effects and security breaches), so that reading progress can be accurately measured over time.

2. The EGRA items and tasks had good reliability values and covered the low-to-middle difficulty range. At baseline, the reading scores were relatively low for both grades, and show room for growth.

In addition, histograms and box pots provided evidence that the tool is expected to measure higher levels of reading that are anticipated due to project-led interventions. Therefore, the baseline data indicates that EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size for statistical power, and the geographic coverage for representation across the province. It should be maintained in the midline and endline, i.e., it should keep the same districts and schools, and continue with the same sampling methods for students at the school level.
4. The systems developed for field data collection should be repeated. The different layers of management, coordination, supervision, and quality control contributed to successful planning, implementation, and problem solving. The quality control officers were particularly important in maintaining standards and providing support for the local subcontractors.
5. The data entry process took time to develop, but it eventually proved to be advantageous in terms of having the data entry operators connect to a central server. This facilitated the two rounds of data entry and the reconciliation process. This system should also be repeated in subsequent data entry activities.
6. The analysis should follow the same procedures, with calculations of reliability, difficulty, task percent-correct scores, summary scores, and fluency (timed) task scores. The baseline, midline and endline scores should be comparable, so that improvements in students' reading can be accurately examined.
7. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.
8. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires to collect data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

In general, the KP baseline was successful in providing accurate data on which to base decisions for implementation of the PRP interventions, and also for tracking student reading progress over time. It provides a solid foundation for the midline and endline assessments.

ANNEXES

Annexes 1 to 4 provide additional information on the EGRA baseline. Specifically, the annexes have the following:

Annex 1 gives complete item statistics – p-values (the difficulty of the items) and item-total correlations (the quality of the items) by grade – for the items associated with the various tasks. These are more detailed than the task statistics presented in Chapter 3 of the report. Measurement specialists often request these kinds of item statistics for the purposes of quality control, analysis, and test equating.

Annex 2 provides box plots for the fluency tasks. The box plots are more task-specific than the overall score distributions (histograms) presented in the report. They show the median (middle score), the range (highest and lowest scores), and the distribution of scores (by quartiles) for each task. The task-specific distributions are useful to EGRA specialists who place emphasis on the fluency tasks.

Annex 3 gives two examples of categorizing passage reading fluency scores using performance levels. The categorizations – along with raw scores and scale scores -- are often used to interpret test scores. The first example combines reading speed with comprehension, while the second example only uses reading speed. Each example uses a set of cut-scores for placing the students into performance categories.

Annex 4 provides detailed information on the second example, with results for each category of fluency and each level of comprehension. These data can be used as evidence on the reliability of using a combined measure of fluency and comprehension for setting performance cut-scores. The validity of combining these scores is more of an issue for reading experts.

Annex 1: Complete Item Statistics by Grade

Table A1 presents item statistics for the untimed tasks, each of which have multiple items. For instance, task 1 (orientation to print) has item statistics for its five items,(Q1 to Q5). Note that the timed tasks are lists of letters, sounds, and words, i.e., not items, so it is not necessary to calculate item statistics for them.

Previously, we presented task statistics (Chapter 3, Table 8) with explanations of how they are calculated. These item statistics are calculated in the same way. They show the difficulty and quality of the items. Recall that when constructing a test, we strive for tasks and items that have difficulty values (p-values) that are spread across the range from about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulty values ranged from 0.03 to 0.81 for grade 3 and 0.06 to 0.85 for grade 5, indicating a strong range of item difficulties. A total of 22 and 21 items for grades 3 and 5 respectively out of the 23 items per grade had item-total correlations of at least 0.20, indicating high quality items.

TABLE A1: COMPLETE ITEM STATISTICS BY GRADE

Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.81	0.31	0.85	0.26
	Q2	0.67	0.24	0.67	0.20
	Q3	0.59	0.26	0.51	0.14
	Q4	0.10	0.04	0.22	0.08
	Q5	0.59	0.22	0.72	0.22
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.54	0.38	0.65	0.43
	Q2	0.33	0.43	0.45	0.54
	Q3	0.30	0.34	0.40	0.39
	Q4	0.28	0.33	0.36	0.40
	Q5	0.32	0.35	0.45	0.44
	Q6	0.47	0.35	0.58	0.45
	Q7	0.19	0.32	0.26	0.43
	Q8	0.26	0.37	0.36	0.45
	Q9	0.22	0.35	0.33	0.46
	Q10	0.51	0.34	0.59	0.45
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.15	0.47	0.30	0.52
	Q2	0.06	0.37	0.14	0.45
	Q3	0.03	0.26	0.09	0.36
	Q4	0.22	0.59	0.42	0.61
	Q5	0.18	0.63	0.41	0.61
8. Listening comprehension (untimed)	Q1	0.16	0.34	0.29	0.40
	Q2	0.03	0.24	0.06	0.22
	Q3	0.32	0.35	0.55	0.37

Annex 2: Box Plots for Phonics and Reading-rate Fluency Tasks

EGRA places a high emphasis on fluency (timed) tasks. In addition to the descriptive statistics in Table 9 (percent correct scores) and Table 14 (fluency task means), we show box plots for the different fluency tasks. Widely used since their development in the 1960s, box plots are a convenient way for graphically presenting numerical data.

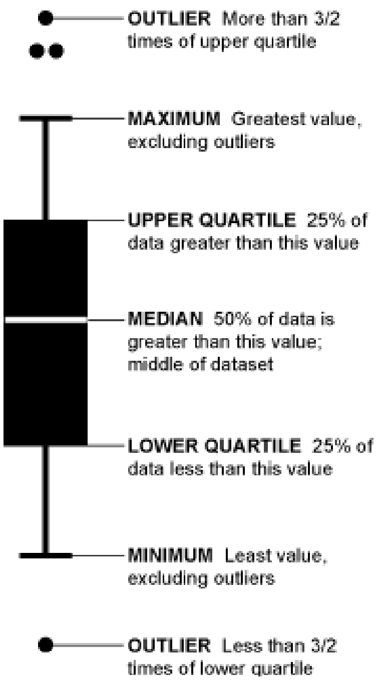
Box plots have two characteristics: 1) central tendency (i.e., the median, or the middle score in the data) and 2) variation (i.e., the range, with scores grouped by quartile). The boxes (which are actually rectangles) represent the two middle quartiles of the scores and the "whiskers" represent the upper and lower quartiles. The small circles on the ends of the whiskers represent outliers. The figure below provides a more detailed explanation for interpreting box plots.

FIGURE A1: UNDERSTANDING BOXPLOTS

Median
The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Upper quartile
Seventy-five percent of the scores fall below the upper quartile.

Lower quartile
Twenty-five percent of scores fall below the lower quartile.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

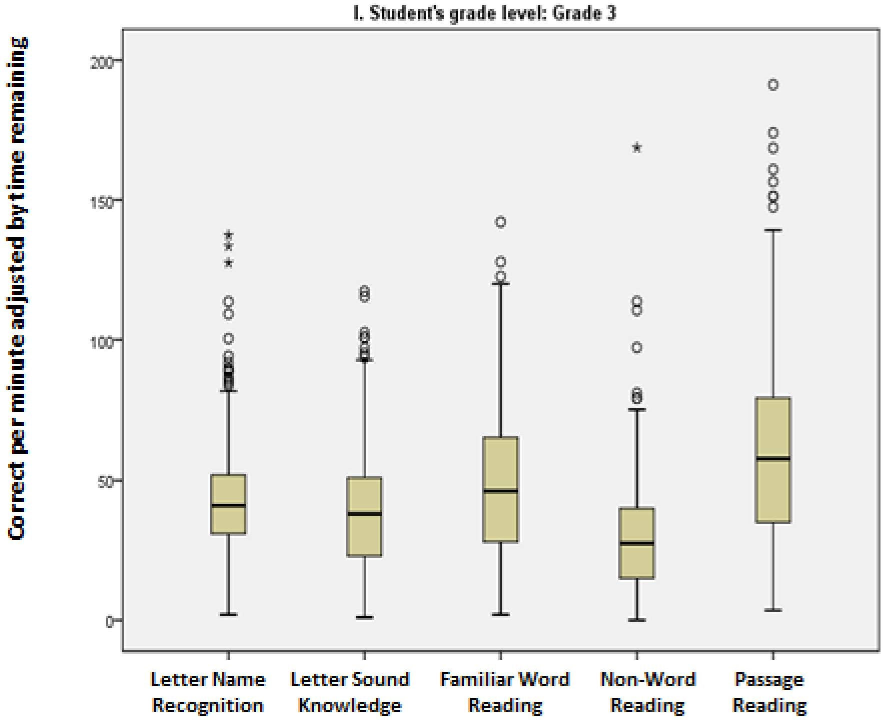
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

Grade 3

For grade 3, the central tendency (i.e., the median speed, or the line in the middle) for each of the tasks ranged from about 30 (non-word reading) to about 60 (passage reading) items per minute. It shows that the students had more fluency reading connected words than conducting grapheme-morpheme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 80 (non-word reading) to about 130 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

FIGURE A2: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 3



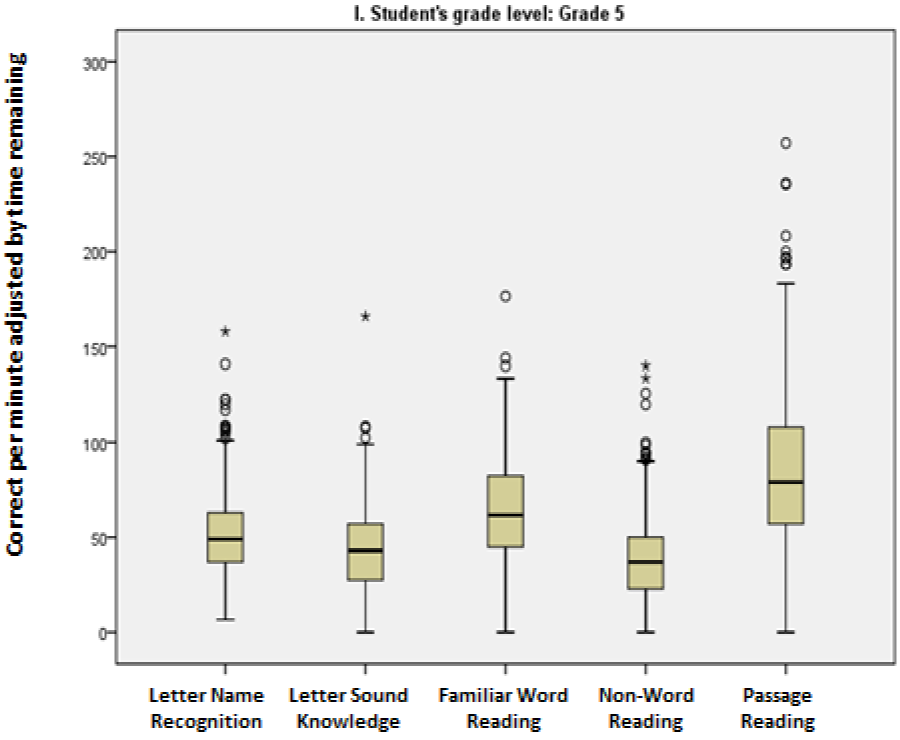
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 40 (non-word reading) to about 70 (passage reading) items per minute. It shows that the students had more fluency with reading connected words than with grapheme-morpheme correspondence.

The variation (range of scores) for each of the tasks varied from about 90 (non-word reading) to about 180 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

Note also that the medians and the ranges increased from grade 3 to grade 5 for all fluency tasks. Many students are becoming more fluent readers at grade 5, but there are also those students who are either non-readers or very low readers. These children lack of knowledge of letter names, sight words, connected text, and (especially) phonics.

FIGURE A3: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 5



Annex 3: Examples of Fluency Score Threshold Calculations

There are different ways of interpreting test scores. Three of the main ways are 1) raw scores (e.g., number correct), 2) scale scores (e.g., percent correct), and 3) percentile scores (e.g., rank in relation to other students). In the report, we presented scores in terms of number correct (for the fluency tasks) and percent correct (for all tasks). We could also calculate the percentile scores for each student, though this is not normally done with EGRA. Note that these kinds of calculations do not change or affect the actual results, but they do involve issues of interpretability.

A fourth main way of interpreting scores is through performance categories, e.g., low, middle, and high. This requires setting cut-scores, or thresholds, to separate the student scores into categories, e.g., two cut-scores lead to three performance categories. The following analysis shows two examples of calculating thresholds for passage reading scores (CWPM), which allows us to place the student scores into different performance categories. Note that performance categories are often accompanied by performance level descriptors (PLDs), which give a text-based explanation of the meaning of the scores in each category. We have not developed PLDs for these examples since 1) the threshold setting is at a preliminary stage and 2) reading specialists with knowledge of local curricula and context generally develop the PLDs.

Fluency using an 80 percent comprehension threshold

In the first example, we used a method that has been suggested by some EGRA specialists. It involves calculating the mean reading speed associated with 80 percent comprehension for those that can read at least one word correctly and then applying it as a fluent cut-score. In other words, the mean reading speed for these students signifies whether the students are fluent readers through using both passage reading speed *and* comprehension in the calculation; the fluent cut-score separates the fluent readers from the non-fluent readers. To establish a second threshold, we again followed the suggested method and used the lowest level of reading (1 CWPM) as the non-fluent cut-score. The two cut-scores resulted in three performance levels: non-readers (low), non-fluent readers (middle), and fluent readers (high).

At grade 3, the mean reading speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 81.0 (or 81). With this method, 81 CWPM becomes a threshold for grade 3 students who are proficient at passage reading *and* comprehension.
At grade 5, the mean speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 108.7 (rounded to 109). Then 109 CWPM becomes a threshold for grade 5 students who are proficient at passage reading and comprehension.

The definitions of the three categories in terms of CWPM and the percentages of grades 3 and 5 students in the categories are shown in Table A2 below.

TABLE A2: THRESHOLDS FOR CWPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	CWPM	% of Students	CWPM	% of Students
Non-Reader	0	23.1%	0	9.9%
Non-Fluent Reader	1 to 80	65.1%	1 to 108	71.8%
Fluent Reader	81 and above	11.8%	109 and above	18.3%
Total	--	100.0%	--	100.0%

Note that the majority of the students are in the middle category at each grade level. This is due the large range of scores for this category, i.e., from the students who score just above non-readers to those who score just below fluent readers are in the non-fluent reader (middle) category.

Fluency using fixed interval thresholds

In the second example, we used fixed intervals of CWPM for the performance levels. This reduced the problem of having a large range of students in the middle category by creating early reader and intermediate reader categories. It also follows common practice when setting performance categories of having between three and five levels for student scores. We used an interval of 40 CWPM to produce five performance levels, along with a category for the non-readers. The five levels were: non-readers (0 CWPM); early readers (1-40 CWPM); intermediate readers (41-80 CWPM); fluent readers (81-120 CWPM); and advanced readers (121 and above CWPM).

TABLE A3: THRESHOLDS FOR CWPM WITH FIXED INTERVALS

Category (Performance Level)	CWPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	23.1%	9.9%
Early Reader	1 to 40	37.7%	17.2%
Intermediate Reader	41 to 80	27.3%	34%
Fluent Reader	81 to 120	10.4%	27.1%
Advanced Reader	121 and above	1.4%	11.8%
Total	--	100.0%	100.0%

At both grades 3 and 5, the fixed interval method allowed for more distribution of the scores across the categories. We can also see a shift in percentages of students in each category from grade 3 to grade 5; the performance categories allow for a score interpretation showing that students are improving across the grade levels, with more scores in the lower categories at grade 3 and more scores in the higher categories at grade 5.

Remarks

While it is possible to use such percentages to set cut-scores for interpretation purposes at the baseline, midline and endline, this analysis should be taken as preliminary. For instance, more well-known and accepted method of setting thresholds – which is commonly called "standard setting" by measurement specialists – involve holding a workshop with local reading experts to set the cut-scores according to the experts' conceptions of what students should know and be able to do in order to be classified into a performance category. There are several well-known methods, e.g., Angoff and Bookmark, which have been judged as valid and reliable for this purpose.[4] Further discussions on setting thresholds involving local reading experts are recommended.

[4] References include: Zieky, M. & Perie, M. (2006). *A primer on setting cut-scores on tests of educational achievement*. Princeton, New Jersey: Educational Testing Service; Cizek, G. (1996). *Standard-setting guidelines*. Educational Measurement: Issues and Practices, Spring 1996, p. 13-21; Cizek, G., Bunch, M., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. Educational Measurement: Issues and Practices, Winter 2004.

Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals

In this last annex, we provide more information on the relationship between reading fluency (speed) and comprehension using information from the fixed interval method. While the data show a positive relationship between speed and comprehension, there are sizeable numbers of "fluent" readers with little comprehension. Our conclusion is that setting a cut-score using a less than reliable indicator, such as the mean speed of students with 80 percent comprehension (i.e., using *both* speed and comprehension), can be problematic. The result is categorizing some students as fluent readers who in fact, according to the definition, are not, i.e., they have high reading speed but low comprehension. It may be better to set thresholds based solely on a single indicator – reading speed – rather than mixing it with comprehension.

The figures and tables below (Tables A4-A5 and Figures A4-A5) expand on the data in Table A3. They show the results for reading fluency (in terms of speed) by comprehension level for grades 3 and 5. We used the categories based on intervals of 40 CWPM, along with a category for the CWPM non-readers (0 CWPM). Comprehension levels were calculated in terms of percent correct scores (e.g., 20 percent is the same as correctly answering one question out of five total questions). For instance, at grade 3, 100 percent of the non-readers have 0 percent comprehension and 22 percent of the advanced readers have 80 percent comprehension.

TABLE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

Category	CWPM	% of Students by Comprehension Level						
(Performance Level)		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	85%	12%	2%	0%	0%	0%	100%
Intermediate Reader	41 to 80	40%	20%	22%	12%	5%	1%	100%
Fluent Reader	81 to 120	22%	14%	25%	29%	9%	1%	100%
Advanced Reader	121 and above	15%	4%	26%	33%	22%	0%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

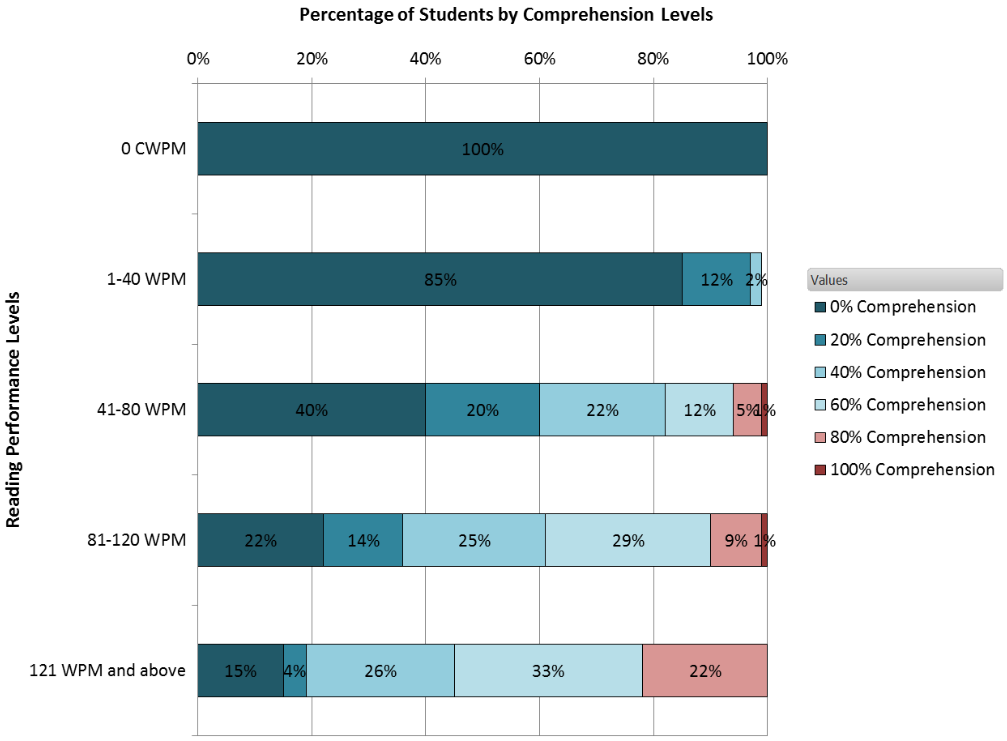
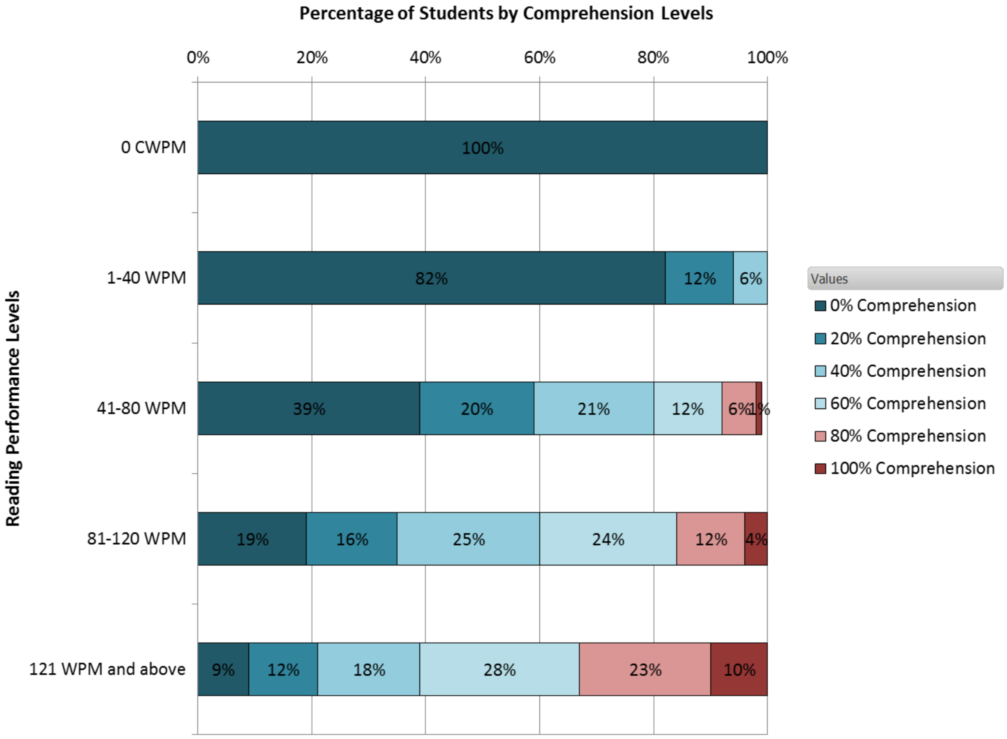


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	CWPM	% of Students by Comprehension Level						
		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	82%	12%	6%	0%	0%	0%	100%
Intermediate Reader	41 to 80	39%	20%	21%	12%	6%	1%	100%
Fluent Reader	81 to 120	19%	16%	25%	24%	12%	4%	100%
Advanced Reader	121 and above	9%	12%	18%	28%	23%	10%	100%

FIGURE A5: GRADE 5 READING FLUENCY AND COMPREHENSION



The main results for the categories of reading speed (from non-readers to advanced readers) in relation to comprehension levels (from 0 percent to 100 percent) for grades 3 and 5 are summarized as follows:

- Non-Readers (0 CWPM) – All of the non-readers had 0 percent comprehension.
- Early Readers (1-40 CWPM) – Most of the early readers (85 percent at grade 3 and 82 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 CWPM) – About two out of every five the intermediate readers (40 percent at grade 3 and 39 percent at grade 5) had 0 percent comprehension. A small minority of them (5 percent at grade 3 and 6 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – About one out of every five fluent readers (22 percent at grade 3 and 19 percent at grade 5 of the) had 0 percent comprehension. Only about one in every ten (9 percent at grade 3 and 12 percent at grade 5) had achieved at least 80 percent comprehension.
- Advanced Readers (121 CWPM and above) – A small percentage of the advanced readers had 0 percent comprehension (15 percent at grade 3 and 9 percent at grade 5). Less than a third (22 percent at grade 3 and 33 percent at grade 5) achieved at least 80 percent comprehension.

The key point from the data is that most of the fluent and advanced readers – at both grade levels – did not reach 80 percent comprehension. Setting a threshold under the assumption that fluent readers (in terms of speed) have a high level of comprehension can be misleading. Conversely, using a single indicator, i.e., reading speed, to set thresholds can be a more reliable way of interpreting the results.